Pdiapredict-Explainer: An Interpretable AI System for Personalized Diabetes Risk Management

Guiyuan Li *

Faculty of Science and Technology, Bejing Normal-Hong Kong Baptist University, Zhuhai, China

* Corresponding Author Email: t330034026@mail.uic.edu.cn

Abstract. This study presents an interpretable LightGBM model designed to predict diabetes risk, developed using the BRFSS2015 dataset, which includes more than 250,000 records. To make which is called the model more understandable and tackle the common "black box" issue, we incorporated feature engineering—such as creating a Metabolic. Risk score—and applied SMOTE to handle class imbalance. The resulting model achieved an AUC-ROC of 0.8264 and was particularly effective at identifying low-risk individuals, with a 98% recall rate, which can help simplify screening efforts. Using SHAP (SHapley Additive exPlanations), we visualized how different factors contribute to each person's risk, including waterfall and decision plots that break down individual cases—like showing how metabolic dysfunction might account for roughly 31% of a person's risk. This approach supports personalized intervention strategies aligned with ADA guidelines. However, the model's ability to catch high-risk cases was limited, with only an 18% recall, emphasizing the difficulty of detecting complex risk profiles without biochemical data like HbA1c. Moving forward, integrating such markers and tracking patients over time with electronic health records could improve performance. Overall, this framework illustrates how explainable AI can build trust in preventive diabetes strategies by making predictions transparent and clinically relevant.

Keywords: Interpretable AI, Diabetes Risk Prediction, Shap Values, Personalized Medicine, Behavioral Intervention.

1. Introduction

Diabetes is a serious, ongoing condition that affects not just the health of individuals, but also their families and communities around the world. The number of people suffering from diabetes was 425 million in 2017. It is estimated that by 2030, there will be an estimated 578 million people suffering from diabetes worldwide, and by 2045, this number may reach around 700 million [1]. It is evident that the growth rate is extremely fast, and people also need to take timely intervention measures. There are three types of diabetes, namely type 1 diabetes (T1D), type 2 diabetes (T2D), and gestational diabetes mellitus (GDM). T1D patients have long relied on insulin injections and are unable to produce enough insulin on their own [2]. T2D happens when people's body cells can't use insulin properly, which actually causes the condition [2]. GDM occurs during pregnancy and is characterized by higher-than-normal blood sugar levels [2]. Most diabetes cannot be cured; it can only be alleviated. To reduce the incidence rate and control the condition, people have begun to use machine models to predict the disease so as to be able to find solutions more quickly. Nevertheless, when using machine learning models, the "black box problem" is prone to exist [3]. When AI systems provide post-diagnosis risk scores without transparent explanations, clinicians cannot verify medical logic or identify biases and thus cannot offer patients more authoritative and accurate treatment methods. Mishra and his companions used the XGBoost model for the prediction of diabetes [4]. While this model achieved the highest scores in precision, recall, and F1, its outputs might be difficult for clinicians and patients to interpret and apply effectively. The lack of transparency can hinder trust and limit its usefulness in a clinical setting [5]. This research was created and improved on the basis of traditional machine learning models. This study went beyond developing a predictive model based on the dataset. By adding a user-friendly front-end and integrating SHAP interpretability analysis, it offers personalized suggestions and feedback mechanisms. These features aim to improve how clinicians and patients understand the patient's condition and explore meaningful options for their health management. Through this model, users can more intuitively see the relevant suggestions for

alleviating their conditions. This has enabled the "black box problem" to be resolved. The interpretability of the model has been enhanced.

2. Methodology

2.1. Data Preprocessing

2.1.1. Data Source:

The Dataset adopted in this study is diabetes_binary_health_indicators_BRFSS2015.csv in the Diabetes Health Indicators Dataset. This dataset is derived from the Behavioral Risk Factor Monitoring System (BRFSS) in 2015 and contains binary labels for health indicators and diabetes (0: no diabetes; 1: Diabetes or prediabetes. The dataset covers information such as demographic characteristics (such as age), clinical indicators (such as BMI, hypertension, high cholesterol), and lifestyle factors (such as physical activity and family history).

2.1.2. Feature Engineering:

The first, the average consumption of Fruits and Veggies was created as the composite feature Healthy_Diet. The composite indicators can better reflect the overall dietary quality than individual features. And delete Fruits and Veggies features will reduce the risk of model overfitting. Then the Metabolic_Risk feature is defined, which is the average of HighBP, HighCol, and BMI greater than 25. The J-shaped risk curve for obesity with a BMI threshold (> 25) is more in line with the pathological mechanism. The next, map the level of physical activity (PhysActivity) to categorical variables, which is low, medium and high. Also, perform individual heat encoding with PhyActivity feature. Finally, delete redundant features, Fruits, Veggies and AnyHealthcare to reduce dimensionality and the complexity of model.

2.1.3. Data Splitting:

The data was divided into a 70% training data, a 10% validation data and 20% testing data through stratified sampling. The training data is applied SMOTE. For the validation data, since the training error of the LightGBM model cannot reliably reflect the generalization performance, it is necessary to monitor the overfitting risk in real time through an independent validation set, which is the core prerequisite for implementing the Early Stopping mechanism. Stratified sampling was used to make sure that the proportion of individuals with diabetes — that is, the prevalence—remained consistent across all subsets.

2.1.4. Class Imbalance Handling:

In view of the class imbalance problem in diabetes data, the synthetic minority class oversampling technique (SMOTE) was applied to the training set. SMOTE achieves a ratio of 1:2 for the minority class to the majority class by generating synthetic samples for the minority class (positive for diabetes). This method expands the decision-making boundary to the majority class area [6]. Early detection of T2D is essential so that interventions can be put in place promptly, helping to prevent missed diagnoses in high-risk populations [7].

2.2. Model Architecture

This study selects the lightweight gradient enhancer (LightGBM) as the core model for diabetes prediction. LightGBM is a fast, efficient and scalable gradient boosting framework developed by Microsoft. Figure 1 and Figure 2 is the flow chart of this project.

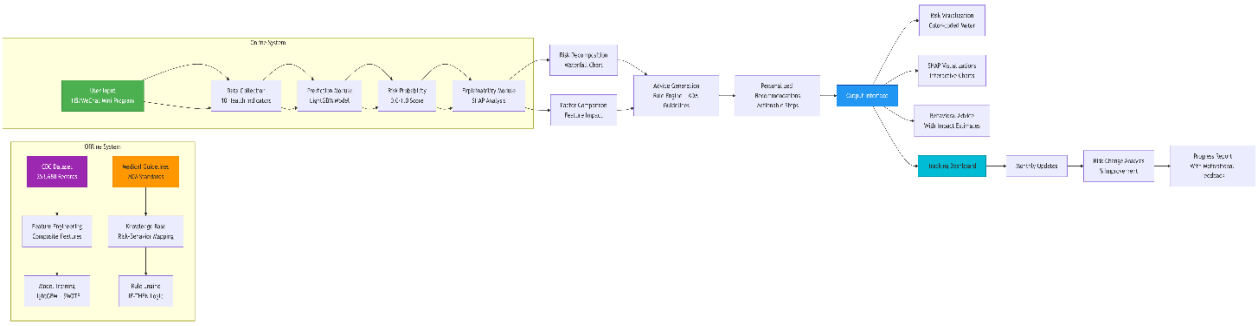


Figure 1: The flow chart 1 (Picture credit: Original).

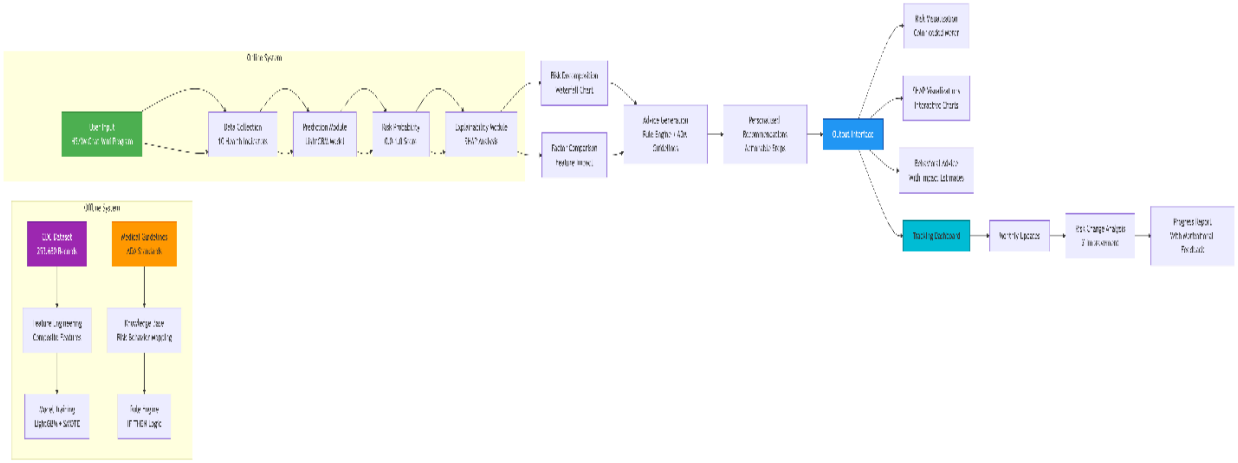


Figure 2: The flow chart 2(Due to the large size of the image, it will be displayed in sections) (Picture credit: Original).

It is widely recognized for its outstanding ability to handle large datasets and its excellent performance in various machine learning competitions. This model is built around three key features. The first is gradient-based unilateral sampling (GOSS), which speeds up training considerably without sacrificing much accuracy by focusing on samples with larger gradients. The second is mutual exclusion feature bundling (EFB), a technique that automatically groups mutually exclusive features to cut down the total number of features and boost processing efficiency. The third approach is to adopt the leaf wise tree growth strategy, generating trees by splitting the leaf nodes with the greatest reduction in loss. Compared with the traditional horizontal growth method, this approach converges faster.

The hyperparameter configuration of the LightGBM model is as follows: The objective function is set to "binary" to meet the requirements of binary classification tasks; The boosting type adopts "gbdt" (Gradient Boosting Decision Tree); The learning rate is set to 0.05 to control the step size of parameter updates. The number of leaf nodes is limited to 31 to prevent overfitting of the model. Set the maximum depth to -1 to allow the tree to grow freely, reaching its optimal size without restrictions. Each child node should have at least 20 samples to ensure there's a solid amount of data in each one. The sample sampling rate is 0.8, and 80% of the samples will be randomly selected for each tree during training. The feature sampling rate is 0.8, and 80% of the features are randomly selected for each tree during training. Simultaneously, L1 regularization (reg_alpha=0.1) and L2 regularization (reg_lambda=0.1) are introduced to further prevent overfitting; The assessment index selects AUC (Area under the curve) as the main evaluation criterion.

During training, the paper implemented an early stopping approach, allowing up to 50 rounds of patience to prevent overfitting. Training was halted when the performance of the validation set did not improve for 50 consecutive rounds to prevent overfitting and determine the optimal number of improvement iterations. A validation set is used throughout training to keep an eye on how the model is performing in real-time. This helps ensure that the model's ability to generalize new data is effectively monitored and adjusted as training progresses. The model is configured to run up to 1,000

training rounds. However, rather than simply sticking to the last iteration, we select the version that achieves the highest AUC score on the validation set. This approach helps ensure we end up with the most effective final model.

2.3. Model Evaluation

The evaluation system takes into account both overall performance and category details. AUC-ROC is adopted as the core indicator to measure the comprehensive ability of the model to distinguish between the two types of samples, namely "risky" and "risk-free". At the same time, the Precision, Recall and F1 score are output through the classification report, with a focus on analyzing the model's performance on the minority class (risky samples). In diabetes screening, guaranteeing that high-risk individuals aren't overlooked is especially important, making the recall rate a key factor to consider. Besides, incorporate SHapley Additive exPlanations (SHAP) values to enhance model interpretability. Using TreeExplainer, we can determine how much each feature contributes to a specific prediction. This analysis produces several visual tools: a waterfall diagram that illustrates the cumulative effect of individual features on the prediction, a decision path diagram that traces the model's reasoning from the baseline to the final output, and a contribution diagram showing whether features have a positive or negative impact and their relative influence. Turning model predictions from obscure "black box" outputs into clear, understandable, and transparent decisions. The calculation of the SHAP value is based on the Shapley value in cooperative game theory. The core idea is: for the prediction result of each sample, it is decomposed into the "baseline value" and the "sum of the SHAP values of each feature" [8].

2.3.1. Waterfall Plot:

The waterfall chart presents the change process of a single sample from the baseline value to the final predicted value in the form of "cumulative superposition", and it is a core tool for understanding the source of individual risk.

2.3.2. Decision Plot:

The decision path diagram presents the complete reasoning path of the model from the baseline value to the predicted value in the form of a "lateral trajectory", and at the same time, it can compare the feature influence differences of multiple samples.

2.3.3. Contribution Plot:

The contribution graph visually presents the SHAP value size and positive or negative direction of each feature in a single sample in a horizontal bar chart. Usually, the 8 features with the greatest influence are selected for display.

SHAP explanations go beyond simply revealing how models generate predictions. They turn raw risk scores into clear, actionable insights through visualizations and detailed analysis, making it easier to understand and act on health information. This approach plays a essential role in supporting personalized decision-making, especially considering diabetes prevention. In addition, we have also added a front-end part to make the result clearer and more straightforward.

3. Results

3.1. Performance of the model

The model demonstrates strong risk discrimination on the test set, achieving an AUC-ROC score of 0.8264. Figure 3 is the Receiver Operating Characteristic Curve.

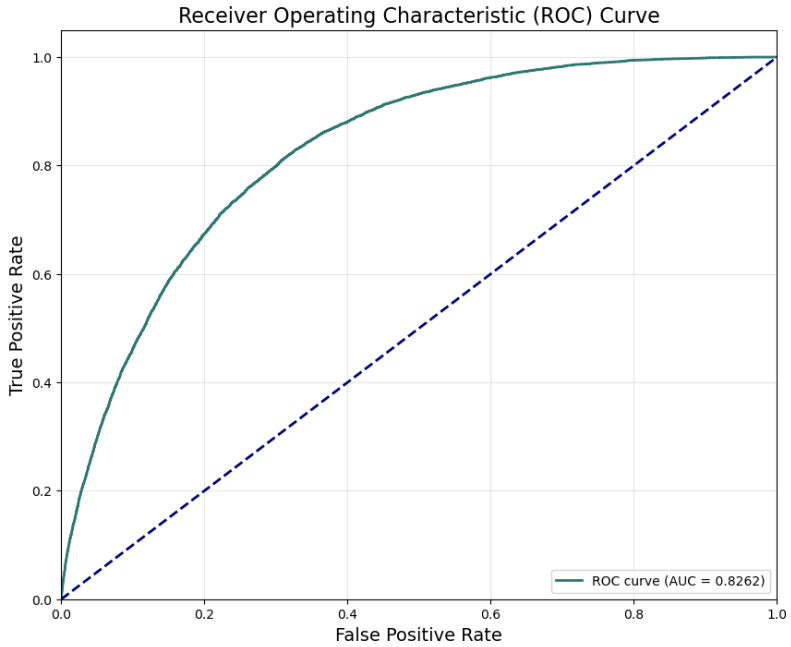


Figure 3. ROC Curve (Picture credit: Original).

This value sits comfortably within the upper-middle range compared to other diabetes risk prediction studies. It suggests that the model is quite capable of differentiating between individuals at higher and lower risk levels and can are a useful reference point for initial clinical screenings.

3.2. Classification indicators

The classification report details how the model performs across various risk categories, emphasizing where it excels and where there's room for improvement. This is the data of this model. Table 1 presents the comparative analysis of the results of different Samples.

Table 1. Analysis and comparison of results from different samples.

	Accuracy	Recall	Precision
Diabetes.binary = 0	88%	98%	93%
Diabetes.binary = 1	Around 55%	18%	27%

The risk-free samples (Diabetes.binary = 0) indicates strong performance in identifying these cases accurately. This suggests that the model is quite effective at correctly ruling out low-risk individuals. As a result, it helps avoid unnecessary follow-up tests and makes better use of medical resources. This result of the samples identified as high-risk (Diabetes.binary = 1) reflects that the model has a relatively high rate of missed diagnosis for high-risk groups. It may be related to two factors. The first is that the characteristics of the minority class (high-risk) in the medical data are highly heterogeneous. Even after SMOTE processing, the synthetic samples still have difficulty fully simulating the complex patterns of real high-risk individuals. The second is that the accuracy of diabetes risk prediction is highly dependent on direct biochemical indicators, but the data of this study only contains behavioral and basic clinical characteristics, lacking such "gold standard" indicators, resulting in the model's limited ability to capture early pathological changes [9].

3.3. Feature Importance

The LightGBM and SHAP analyses indicate that the influence of the main risk factors, from greatest to least, is: Metabolic_Risk, BMI, HighBP, FamilyHistory, which is in line with the conclusions of the clinical studies. Metabolic syndrome, which includes high blood pressure, enhanced cholesterol levels, and obesity—is a important indicator that prediabetes may be developing. Its comprehensive impact is more capable of reflecting the potential problems of the body (such as

insulin resistance) than a single indicator [9]. In the model, the average impact of the metabolic risk index on the risk is + 0.23, and it is the most influential factor. As mentioned by the American Diabetes Association, for every 5 kg/m² increase in BMI, the risk of type 2 diabetes rises by approximately 80%. For individuals with a BMI above 25, the model indicates that this factor increases the risk by an average of 0.18. What's more, as BMI rises, its influence on risk tends to grow, aligning well with clinical observations.

3.4. Explanation SHAP

SHAP effectively emphasizes the key factors influencing each person's risk profile, allowing healthcare professionals to offer targeted intervention recommendations. This capability is particularly evident in two main areas.

3.4.1. The Waterfall chart decomposes risks:

A 55-year-old individual with high metabolic risk and a sedentary lifestyle can see how their overall risk score of 0.72 is determined. The waterfall chart breaks down the calculation, starting from a base risk of 0.15 and adding contributions from various factors: metabolic issues (+0.31), weight (+0.19), and age (+0.08). Using SHAP values helps clearly identify which factors have driven that increased risk, giving the doctor a straightforward way to recommend targeted actions—like boosting physical activity, losing weight, and improving metabolic health. Figure 4 is his waterfall chart.

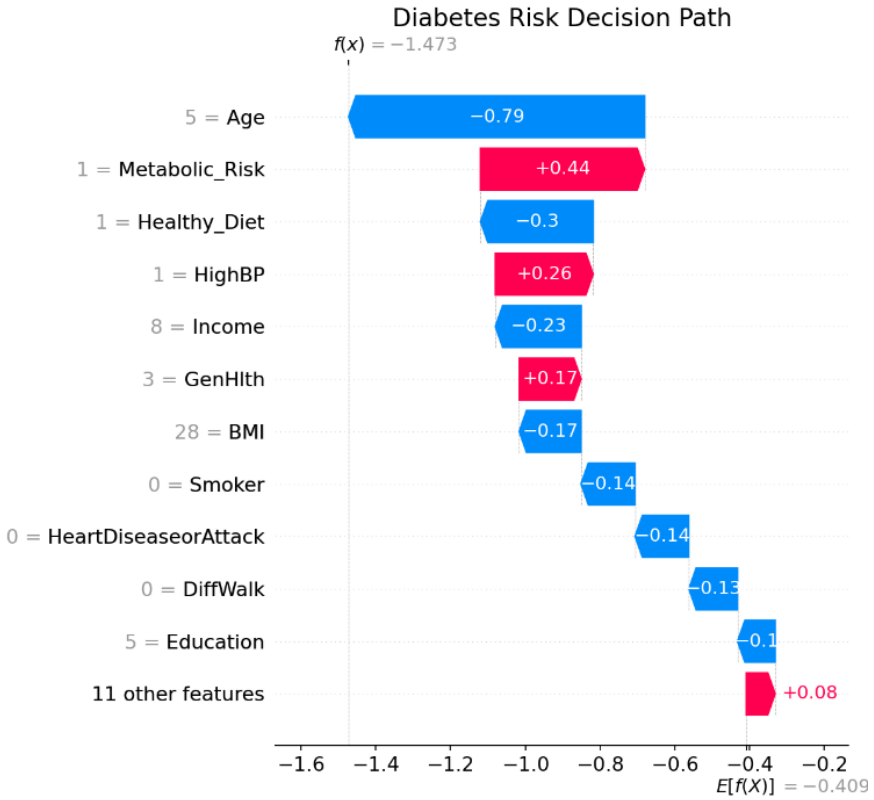


Figure 4: Waterfall Chart (Picture credit: Original).

3.4.2. Compare the risk causes of different people:

Two individuals with similar weight and height showed noticeably different health risks, largely influenced by their activity levels. The person who was more active had a risk score of 0.32, greatly lower than the 0.58 risk seen in the less active individual. The SHAP analysis indicated that physical activity was the biggest factor, accounting for a risk reduction of 0.21. Engaging in about 150 minutes of moderate-intensity exercise each week can cut the chances of developing diabetes by roughly 15 to 20%. This kind of insight can help healthcare providers offer more customized advice based on individual activity habits. Figure 5 shows the comparison of the risk causes of different people [10].

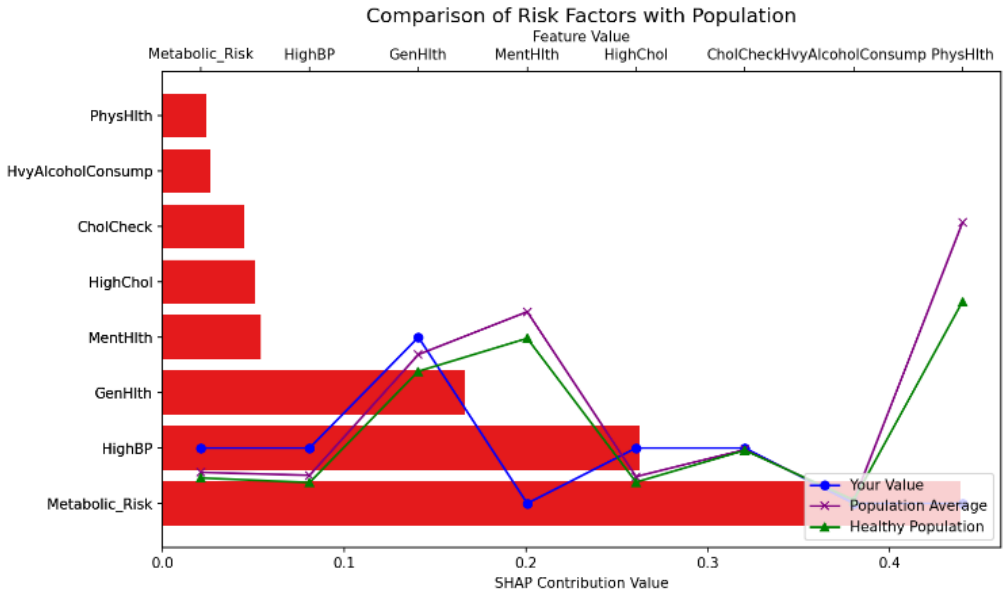


Figure 5: Comparison Chart (Picture credit: Original).

4. Conclusion

This study combined clinical features with machine learning methods to create an interpretable model for assessing diabetes risk. The goal was to develop a practical tool for personalized screening in primary prevention settings. The model achieved an AUC-ROC of 0.8264 on the test set and a recall rate of 98% for identifying low-risk individuals. This suggests it can efficiently exclude healthy people during large-scale screenings, helping to conserve medical resources. The SHAP analysis supported the model's clinical relevance, emphasizing key risk factors such as metabolic risk index, BMI, and hypertension—those that align well with current understanding of diabetes development. Also, breaking down individual feature contributions offers clear targets for personalized intervention strategies.

The model's ability to identify high-risk individuals currently has a recall rate of only 0.18, which raises concerns about missing potential diagnoses in clinical settings. One key reason for this limitation is the absence of direct biochemical indicators—such as blood glucose and insulin levels—in the dataset. Besides, the diverse characteristics of high-risk samples contribute to this challenge. This issue mirrors a common problem in medical AI models, where recognizing minority classes remains difficult. Moving forward, incorporating gold standard diagnostic markers like glycated hemoglobin (HbA1c) could improve early detection of pathological changes. Using ensemble learning to combine multiple models might also boost recognition of complex risk patterns. What's more, integrating longitudinal data from electronic health records (EHR) will help develop energetic risk prediction models that better reflect disease progression over time.

This study confirms that machine learning is both feasible and valuable for assessing diabetes risk. The way we used SHAP interpretation to connect features, risk levels, and interventions offers a clear example of how AI can be trusted in medical applications. Moving forward, we need to improve the model's accuracy in spotting high-risk individuals and validate its effectiveness through real-world clinical trials.

References

[1] Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., ... Williams, R. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045. *Diabetes Research and Clinical Practice*, *157*, 107843. https://doi.org/10.1016/j.diabres.2019.107843

[2] Choubey, S., Agrahari, S., Shaw, A., Dhar, S., Sarma, R. R., Singh, S. K., Das, P., & Saha, B. (2023). Diabetes prediction using ML. *International Journal for Research in Applied Science & Engineering Technology*, *11*(6), 4209–4212. https://doi.org/10.22214/jiraset.2023.54415

[3] Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, *20*(1), 310. https://doi.org/10.1186/s12911-020-01332-6

[4] Mishra, V. T., Mohan, R., Sreedharan, J. (2025). Prediction of 30-day readmission in diabetes management using Machine learning. *Computers in Biology and Medicine*, Volume 108, Article 110616. https://doi.org/10.1016/j.compbiomed.2025.110616

[5] Rudin, C. (2019). Stop explaining black box machine learning models. Nature Machine Intelligence, *1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

[6] Arrieta, A. B., Díaz-Rodríguez, N., Bennetot, A., Tabik, S., Barbado, A., García, S., ... Herrera, F. (2020). Explainable AI in healthcare. *Artificial Intelligence in Medicine*, *110*, 101964. https://doi.org/10.1016/j.artmed.2020.101964.

[7] American Diabetes Association. (2022). *American Diabetes Association releases 2023 standards of care in diabetes to guide prevention, diagnosis, and treatment for people living with diabetes* [Press release]. https://diabetes.org/newsroom/american-diabetes-association-2023-standards-care-diabetes-guide-for-prevention-diagnosis-treatment-people-living-with-diabetes

[8] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[9] American Diabetes Association Professional Practice Committee. (2024). Introduction and Methodology: Standards of Care in Diabetes—2024. *Diabetes Care*, Volume 47(Supplement_1), S1–S4. https://doi.org/10.2337/dc24-SINT

[10] American Diabetes Association. (2022). *Standards of medical care in diabetes—2022*. American Diabetes Association. https://doi.org/10.2337/dc22-Sint.